they will detect no changes. But if required to guess which pictures do involve a change, they might well do better, even much better, than chance, in spite of their utter inability to say what these changes might be. (This experiment is currently under development in Rensink's lab.) If subjects can make good forced-choice guesses, this would conclusively show that some information was preserved from moment to moment, that there was *some* nonephemeral "registration" after all. This would not serve Dretske's purposes, however, by giving him a "difference which makes a difference" on which to hang nonepistemic seeing *in conscious experience*, since this performance on forced-choice guessing is precisely the evidence standardly relied on to demonstrate *unconscious* information preservation in blindsight. I doubt that Dretske would want blindsight to count as nonepistemic seeing.

So, to revert to the confrontation with which we began, Dretske noted what he takes to be a clear mistake of fact in my theory of consciousness. I say the detail only *seems* to be "in there" and he disagrees. I agree that it is in the eye (focused on the retina), but that is surely not enough, for in that sense, the detail is also in the *camera obscura*. Most of this detail is not—and cannot be—picked up *at all*, but some of it is. The few details that are picked up are picked up *by being identified* or categorized in some fashion—if only as blobs worthy of further consideration, as Treisman's experiments show.[43] It does indeed *seem* as if all the details are "in here" in some stronger sense—a difference that makes a difference—but that is an illusion.

III. TRUTH-MAKERS, COW-SHARKS, AND LECTERNS

McLaughlin and O'Leary-Hawthorne have succeeded where others have tried and failed. They have obliged me to respond, in detail worthy of their challenge, to the question: Why don't I take Swampman, Blockheads, and their friends seriously? They have obliged me by writing an exemplary essay, fair, patient, and reasonable, setting out the problems with my view as they (and many other philosophers) see them. They provide a compelling exhibition of something philosophers should more often strive for: A consideration of ideas that transcends questions of who said what, when. Which of the many variations of the ideas they consider is mine? Which did I mean? It doesn't much matter, since they canvass all the possibilities, and try to show which is the best—given my purposes—and why. If I didn't say or mean that, I should have. Or so they claim, with supporting reasons. First let me confirm a suspicion that they hint at occasionally: I have not thought that such fanatic attention to precise formulations was work worth doing; I still think that this is largely make-work, but there are many, apparently, who think I am wrong, and I owe them, in my response to this challenge, a proper reply.

It cannot have escaped philosophers' attention that our fellow academics in other fields—especially in the sciences—often have difficulty suppressing their incredulous amusement when such topics as Twin Earth, Swampman, and Blockheads are posed for apparently serious consideration. Are the scientists just being philistines, betraying their tin ears for the subtleties of philosophical investigation, or have the philosophers who indulge in these exercises lost their grip on reality?

These bizarre examples all attempt to prove one "conceptual" point or another by deliberately reducing something underappreciated to zero, so that What Really Counts can shine through. Blockheads hold peripheral behavior constant and reduce internal structural details (and—what comes to the same thing—intervening internal processes) close to zero, and provoke the intuition that then there would be no mind there; internal structure Really Counts. Manthra is more or less the mirror image; it keeps internal processes constant and reduces control of peripheral behavior to zero, showing, presumably, that external behavior Really Doesn't Count. Swampman keeps both future peripheral dispositions and internal states constant and reduces "history" to zero. Twin Earth sets internal similarity to maximum, so that external context can be demonstrated to be responsible for whatever our intuitions tell us. Thus these thought experiments mimic empirical experiments in their design, attempting to isolate a crucial interaction between variables by holding other variables constant. In the past I have often noted that a problem with such experiments is that the dependent variable is "intuition"—they are intuition pumps—and the contribution of imagination in the generation of intuitions is harder to control than philosophers have usually acknowledged.

But there is also a deeper problem with them. It is child's play to dream up further such examples to "prove" further conceptual points. Suppose a cow gave birth to something that was atom-for-atom indiscernible from a shark. Would it be a shark? What is the truth-maker for sharkhood? If you posed that question to a biologist, the charitable reaction would be that you were making a labored attempt at a joke. Suppose an evil demon could make water turn solid at room temperature by smiling at it; would demon-water be ice? Too silly a hypothesis to deserve a response. All such intuition pumps depend on the distinction spelled out by McLaughlin and O'Leary-Hawthorne between "conceptual" and "reductive" answers to the big questions. What I hadn't sufficiently appreciated in my earlier forthright response to Jackson is that when one says that the truth-maker question requires a conceptual answer, one means an answer that holds not just in our world, or all nomologically possible worlds, but in all logically possible worlds.[44] Smiling demons, cow-sharks, Blockheads, and Swampmen are all, some philosophers think, logically possible, even if they are not nomologically possible, and these philosophers think this is important. I do not. Why should

the truth-maker question cast its net this wide? Because, I gather, otherwise its answer doesn't tell us about the *essence* of the topic in question. But who believes in real essences of this sort nowadays? Not I.

Consider the fate of "logical behaviorism" with regard to magnets. Here are two candidate answers to the question of what the truth-maker is for magnets: (*a*) all magnets are things that attract iron, and (*b*) all magnets are things that have a certain internal structure (call it M-alignment). Was the old, behavioral criterion (*a*) eventually superseded by the new, internal structure criterion (*b*), or did the latter merely reductively explain the former? To find out, we must imagine posing scientists the following Swampman-style questions. Suppose you discovered a thing that attracted iron but was not M-aligned (like standard magnets). Would you call it a magnet? Or: Suppose you discovered a thing that was M-aligned but did not attract iron. Would you call it a magnet? The physicists would reply that if they were confronted with either of these imaginary objects, they would have much more important things to worry about than what to call them.[45] Their whole scientific picture depends on there being a deep regularity between the alignment of atomic dipoles in magnetic domains and iron-attraction, and the "fact" that it is logically possible to break this regularity is of vanishing interest to them. If they are "logical behaviorists" about magnets, this is no doubt due to William Gilbert's early phenomenological work in the seventeenth century, which established the historical priority, if nothing else, for the classification of magnets by what they do, not what they have inside. (He built upon, and improved, the folk physics of magnets, in short.) What is of interest, however, is the real covariance of "structural" and "behavioral" factors—and if they find violations of the regularities, they adjust their science accordingly, letting the terms fall where they may. Nominal essences are all the essences that science needs, and some are better than others, because they capture more regularity in nature.

In "Do Animals have Beliefs?" I say, commenting on a point of agreement between Fodor and me:

> We both agree that a brain filled with sawdust or jello could not sustain beliefs. There has to be structure; there have to be elements of plasticity that can go into different states and thereby secure one revision or another of the contents of the agent's beliefs.[46]

Doesn't this passage concede everything McLaughlin and O'Leary-Hawthorne have been pressing on me? When I say "could not" and "have to," am I speaking of "conceptual" or "nomological" necessities? I am speaking of serious necessities. If I ever encounter a plausible believer-candidate that violates them, what to call it will be the least of my worries, since my whole theory of mind will be sunk.

So why do I lean towards "logical behaviorism" and away from the specifics of internal activity and structure that McLaughlin and O'Leary-Hawthorne go to such lengths to highlight? For the reasons that **Lynne Rudder Baker** explains so well in her essay. Like Gilbert, I start with folk theory, which is remarkably robust in the case of folk psychology. It is a discovered fact, already well confirmed, that "peripheral narrow behavior" of the sorts commonly observed by everyday folk, is readily predicted and explained by folk psychology. Thus the order of explanation is from outer to inner, not vice versa. We want a theory of the innards that can account for *all that regularity*. It *might* have gone otherwise; it is logically possible, I suppose, that we could have found "belief boxes" in people's heads that causally explained their behavior and well-nigh identical "belief boxes" in the cores of redwood trees that were entirely inert. We would then have put a premium on explaining *that* regularity of internal structure, and let the differences in behavioral consequences tag along behind. But we didn't find any such thing. It is not just logically possible but already demonstrated that there are in fact many internally different ways of skinning the behavioral cat, while it is at best logically possible, and Vanishingly[47] unlikely, that we will ever encounter Manthra, or anything else that is an internal twin lacking the behavioral prowess.[48]

This all depends, of course, on how closely we look at the innards for signs of similarity. How different do internal ways have to be to count as different? McLaughlin and O'Leary-Hawthorne see a contradiction between my various positions on behaviorism, and I guess they are right. I should have *explained* why I thought that the difference between molecular and molar behaviorism didn't amount to anything important, rather than burking the distinction altogether. Of all the molecular differences that there might be, the only ones that would make a difference *to psychology* (as ordinarily understood) would be those that made a difference to the "peripheral narrow behavior" that is predicted and explained by folk psychology.

Consider: Tweedledum and Tweedledee both hear a joke and both laugh; both also *would* laugh at various other jokes, would find others unfunny, etc. Nevertheless their overall joke-getting machinery has some differences—differences that would never show up in any peripheral behavior. These differences are clearly (I would think) below the level of psychology. In particular, these differences would not license a different attribution of belief. Start with what is probably a safe limiting case. Tweedledum's brain makes somewhat different use of potassium in its regulation of axonal transmission than Tweedledee's brain does. Otherwise, their brains always "do the same thing"—they are not quite molecular behavioral twins, but pretty close. Though not molecular behavioral twins, they are nevertheless psychological twins, for the differences are just too fine-grained to show up in interesting psychological differences—such as different belief

attributions on anybody's story of belief attributions. Suppose next a much larger-scale difference: Tweedledum and Tweedledee have entirely different subcognitive systems of face recognition; one relies on a sort of feature-detection checklist, and the other on some global, holistic constraint-satisfaction scheme.[49] Now, we may suppose, they do exhibit different *psychological* profiles (at least in relatively abnormal circumstances): Under experimental conditions, one can readily recognize faces that the other cannot, and hence they will not share just the same beliefs. If, contrary to plausibility, their radically different "face-recognition modules" had exactly the same competence under all experimental conditions, we would see the difference as an interesting physiological difference but too fine-grained to "count" as psychology. But where we draw the line is not a big deal, one way or the other.

Some philosophers may still think that in spite of all this, Blockheads illustrate an important principle, so before taking my leave of Blockheads I cannot resist pointing out that the "principle" relied upon by Block in his original thought experiment is mistaken in any case. One of the most intelligent things any thinking agent can do is plan ahead, engaging in what we might call *temporally distal self-control*. Anticipating that when push comes to shove at some later time, it may be difficult or impossible to Do the Right Thing—to figure out and execute the rationally optimal response to current circumstances—the wise agent arranges to tie his hands a little and cede temporally local control to a policy figured out long ago, in cool, dilatory reflections "off line." Dieters, knowing their urges, arrange to locate themselves in places bereft of snacks, and when they later act in an environment that does not include "shall I have a snack?" as a live option, this is a feature of the environment for which they themselves are responsible, as a result of their own earlier intentional actions, not a mere external constraint on performance. The practical navigator, John Stuart Mill reminds us (in *Utilitarianism*), goes to sea with the hard problems of spherical trigonometry and celestial motion precomputed, their answers neatly stored in a rather large (but portable) look-up table. It is thus no sign of mindlessness, but rather of foresight, if we encounter the navigator mechanically determining his position by looking up the answers, swiftly, in a book.

We think of Oscar Wilde as a great wit. It would no doubt diminish his reputation considerably if we learned that he lay awake most nights for hours, obsessively musing "Now what would I reply if somebody asked me. . . , and what might my pithy opinion be of . . . . ?" Suppose we learned that he patiently worked out, and polished, several dozen bon mots per night, ingeniously indexing them in his memory, all freshly canned and, if the occasion should arise, ready to unleash without skipping a beat—for brevity is indeed the soul of wit. They would still be *his* bon mots, and their funniness would depend, as the Polish comedian said, on timing. Timing is important

for almost any intelligent act—which is why it is intelligent to anticipate and pre-solve problems wherever possible. Wilde's brute force witticism-production scheme might disappoint us, but in fact it draws attention to the fact that all intelligent response depends on costly "R and D", and it doesn't make much difference how the work is distributed in time so long as the effects are timely.

So, contrary to Block's guiding intuition, discovering that some Turing Test contestant was (one way or another) looking up the responses in a giant look-up table should not at all rule out the hypothesis that this was the manifestation of an intelligent agent at work. Local inspection would perhaps often leave us in doubt about who the intelligent agent was (or who they were), but we should have no doubt at all that the witticisms on the transcripts were the product of intelligent design, responsive to the meaning of the inputs, and just temporally removed by being solved in the hypothetical. Intelligent design is the only way witticisms can be made. Am I saying it is actually *logically necessary* that any such giant look-up table of clever responses would have such an etiological history? Heavens no. A cow-shark could give birth to one. But in our world, the only way anything will ever pass the Turing Test is by being an intelligent, thinking agent.

Mark Richard provides a close encounter that is long overdue, confronting my "pretty pernicious instrumentalism" with a relentless challenge from one of those who think that the way to make a proper theory of belief is to construct and defend formal definitions of its terms. I turned my back on the efforts of the Content Mafia (otherwise known as the Propositional Attitude Task Force) in 1982, after publishing "Beyond Belief," in which I gave my reasons for rejecting their methodology and enabling assumptions. The tradition has continued in force without my blessing, of course, and few participants have felt the need either to respond to my criticisms or to show how their way of philosophizing shows what is wrong with mine, so it is high time to see how the scales balance today.

Richard offers a three-pronged attack on my account of believers as intentional systems: I cannot solve the lectern problem, he claims, and two avenues which might seem to offer escape hatches for me, Stich's attempt to distinguish beliefs from subdoxastic states via a condition of "inferential promiscuity," and Evans' Generality Constraint, turn out to be flawed. If, as Richard notes, I can't adequately answer the question "What isn't a belief?" I can't answer the question "What is a belief?" either. Not a good verdict for a theory of belief.

I claim to solve the lectern problem, as Richard observes, by showing how, when predicting lectern "behavior," the intentional stance gives a predictor no purchase over using the physical stance. "That no *single event* is unpredictable from the Laplacean perspective does not imply that every *behavioral pattern* is perceptible from the perspective," Richard notes, but

he then tries to shoehorn my position into the assertion that such behavioral patterns can be identified with the "instantiation of a ceteris paribus law of property instantiation by an individual."[50] No, the impressive patterns cannot be reported in *single* generalizations of the sort Richard illustrates; the patterns that inspire adoption of the intentional stance consist in the success (with *n*% noise) of a myriad of such ceteris paribus generalizations. No one predictive success counts for much at all—witness the lectern's readily predicted null behavior.

The predictive power of the intentional stance does not derive from our having induced kazillions of psychological "laws" which we are reminded of whenever we see their antecedent conditions being satisfied. Where would all these "laws" come from? We surely aren't taught them by the score. Rather, we effortlessly *generate* our predictions from an appreciation of the underlying normative principle of intentional stance prediction—rationality. What we need the strategy to explain is our power to generate these predictions—describe these patterns—ad lib and ad hoc in any number we wish, and find the vast majority of them to be predictive way better than chance. That's why I spoke of intentional systems whose behavior is "voluminously" predictable, a theme Richard notes in passing, but underestimates.

Richard then goes on to interpret my claim about the ineliminability of the intentional characterization of people (and other true believers) as the claim that the presumptive intentional laws governing lecterns, unlike the intentional laws governing believers, have equally predictive—indeed coextensive—"physical equivalents," and hence are eliminable. This misconstrues my case. I am quite willing to grant that some unimaginably long but finite disjunction of physically characterized conditions exhausts by brute force the entire predictive power of any intentional stance prediction whatsoever. So what? Such claims are not interesting; the same move could be used to strike down every biological category (for instance), since the Heat Death of the Universe, if nothing else, guarantees that "there is" a huge but finite disjunction of predicates constructible by Boolean means from terms drawn strictly from subatomic physics that would be exactly as predictive as "is a herbivore" or "is hemoglobin" or "reproduces asexually." And yet the patterns referred to by these biological terms are perfectly real.

The guaranteed existence of such an unwieldy predicate doesn't diminish the actual value, for purposes of prediction, of an intentionalistic predication, and it doesn't explain what an intentionalistic claim explains. I have improved on my Martian predictor example in *Darwin's Dangerous Idea*, in the riddle of the two black boxes.[51] I show that there are short, readily tested *causal generalizations* whose almost exceptionless truth would be manifest to super-Laplaceans but utterly mysterious and inexplicable by them unless they adopted the intentional stance. The fact that the super-Laplaceans could predict each instance of the generalization—could

generate, given enough time, every disjunct in the unimaginably long list—would not impress *them*, since they would see clear as day in the totality of their predictions a simple regularity that they could not explain.

Why does Richard go to such pains to translate my thesis into the alien language of "ceteris paribus laws" and then interpret its central claim in terms of the nonexistence of equivalent sentences composed in nonintentional vocabulary? The reason, I think, is methodological: Richard simply cannot use the tools of his trade unless he can first turn the object of scrutiny into such a claim. This methodological imperative comes out more sharply when we turn to his painstakingly constructed arguments against Stich's distinction and Evans' Generality Constraint. Here he helps himself on several occasions to the tempting surmise that there is a language of thought, even though he concedes that this hypothesis may not, in the end, make much sense when applied to human believers. Why does he do this? Because he cannot construct his arguments without it. He is not alone, but his forthrightness on this occasion helps to underscore a point I have made in the past: All the fine arguments spun from the intricate examples posed by the Propositional Attitude Task Force depend on isolating rather special cases of what I have called opinions—linguistically infected states quite distinct from beliefs—and showing how, if these presumably clearly identified propositional attitudes are held constant in imagination, troubles can be raised. Without the language of thought as a crutch to keep the "fatal" examples from toppling over under the weight of their often bizarrely top-heavy loads of specific content, there would be no research program here at all.[52]

> Even if *our* thought is not invariably realized in a linguistic medium, the existence of something whose thought was invariably so realized and in whom we would identify the possession of concepts with lexical mastery, isn't impossible.[53]

Or, rather, such a being better not be impossible, since a cottage industry of philosophical research depends on it. Part of my evidence for this claim is the studied indifference of these philosophers, almost without exception, to the efforts by various researchers in Artificial Intelligence and cognitive psychology to construct and defend models of belief that do utilize something like a language of thought. Is Douglas Lenat's CYC project the sort of entity Richard has in mind? It is a "belief box" containing millions of hand-coded propositions, and insofar as it has any concepts at all, it is in virtue of the "lexical mastery" provisioned by all those carefully wrought definitions, as interanimated by its attached inference engine. Is CYC a believer? The general run of opinion in cognitive science is that CYC is a brave attempt at an impossible project. At the very least, the burden of proof is on those who think that it *is* possible for something we would recognize as a believer or thinker to be composed as Richard assumes. (Of course it is possible to con-

struct large boxes of interanimated sentences—CYC is an actual instance—but few would think that a theory about which sentences appeared where under what conditions in such boxes would be a theory of belief. That is, however, one way of reading the underlying assumption of a language of thought.)

Richard appeals at one point to the supposition "I think in English," and at another point tells us of Jane, who believes "Twain's here" expresses a truth but doesn't realize that Mark Twain is Clemens. Later he alludes to Smith, the poorly integrated bilingual, and finally to Jan, the bilingual Dutchman. He needs these special cases because he has to be able to point to propositions crisply "identified" as only a specific sentence in a particular language can do (one can speak about sets of possible worlds, but the only practical way of saying which set you have in mind is to go piggyback on a specific sentence). For instance, Jan's belief that lions are in zoos has to be identified with a specific sentence in one of Jan's languages of thought, so it can be clung to, as one of Jan's beliefs, in spite of the evidence that Jan is really a bit dense about lions and zoos, so dense that he can't even "think the thought" in his other language of thought. Richard tells us the point of the exercise: "What Jan provides us with is *an example* of a believer for whom the Generality Constraint fails."[54]

The grandfather of all such cases is Kripke's Pierre,[55] who believes—well, what *does* he believe about London? While hundreds of pages have been published about Pierre, I have not bothered adding to them, since the proper response seemed to me to be so obvious.[56] Thanks to Kripke's clear setting out of the conditions under which Pierre fell into his curiously ill-informed state, we know exactly what his state of mind is. What is the problem, then? The problem is saying, formally and without fear of embarrassing contradiction elsewhere, *exactly* what Pierre believes. Which propositions, please, should be inscribed on Pierre's belief list, and how are they to be individuated? Well, it can't be done. That's the point of the Pierre case; it neatly straddles the fence, showing how the normally quite well-behaved conditions on belief pull against each other in abnormal circumstances. What should one do in such a dire circumstance? Chuckle and shrug, and say, "Well, what did you expect? Perfection? Pierre is an imperfect believer, as we all are."

How can I say we know exactly what Pierre's state of mind is while cheerfully admitting that we cannot *say*, exactly, what his state of mind is *in terms of propositional attitudes*? Simple: Propositional-attitude talk is a hugely idealized oversimplification of the messy realities of psychology. Whenever push comes to shove in borderline cases, its demands become unanswerable. That is my pretty pernicious instrumentalism showing, I guess. I don't call my view instrumentalism anymore, but whatever it should be called, my view is that propositional-attitude claims are so idealized that it is often impossible to say which approximation, if any, to use. There is

nothing unprecedented about this: Biologists shrug when asked whether herring gulls and lesser black-backed gulls are truly different species,[57] and electrical engineers are unperturbed when you point out that it is quite possible to take a perfectly good FM tuner and, by making a few minor revisions, turn it into something that is maybe a genuine but lousy FM tuner and maybe not an FM tuner at all. How close to the (ideal) "specs" does something have to be to count as a genuine FM tuner? What if it can receive only one station? What if it tends to receive two stations at a time? What if a cowshark swallows it, and its stomach acid turns it into a television set?

The various predicaments that Richard treats as counterexamples to theories could better be considered to be shortcomings in the particular believers, fallings short from the ideal of inferential promiscuity, or Generality, for instance. (Pierre is a true believer, of course, but a decidedly suboptimal one. Believers aren't *supposed* to get themselves into the sort of epistemic pickle Pierre has blundered into.) Since all believers fall short of the ideal, Stich's useful idea about how to tell beliefs from other "subdoxastic states" should be treated as a desideratum of beliefs, not a litmus test. Then we can see a gradation of cases, from truly embedded or encapsulated subdoxastic states to more and more "movable" and inferentially available states. The question of how, in the species and in the individual, this transition to more and more versatile cognitive states occurs is fast becoming a major theoretical issue in cognitive science.[58] It wisely ignores the question of how to define belief formally. If, on the other hand, you insist on setting up a definition of belief as a set of necessary and sufficient conditions, in the fashion that Richard assumes obligatory, you merely guarantee that there are no true believers, not among ordinary mortals.

So Richard is mistaken in thinking from the outset that I take on the obligation to offer a "principled account of the distinction between (having) propositional attitudes as against (having) psychological states which, though they produce and regulate behavior and can be assigned informational content, are not propositional attitudes."[59] And hence he is mistaken about the role that either Stich's or Evans' claims might play for me, whatever role they have played in the work of others. I myself have always thought that the Generality Constraint nicely captured the ideal—the same ideal Fodor captures by speaking of belief fixation as "Quinian and isotropic." It is an ideal no believer meets but all—all worthy of the name—approximate. One of the best arguments against CYC-style models of belief could in fact be put thus: Since it is at least very hard (and maybe impossible) for them to meet Evans' Generality Constraint, even in approximation, there must be some other way of organizing the innards of a believer that accounts for the fact that believers are in general quite able to honor that constraint.

The trouble with the tools of the trade of the Propositional Attitude Task Force is that they cut too fine! Propositions are abstract objects and

(according to theory) just as distinct and well-behaved as, say, integers. If propositions measure psychological states the way numbers measure physical states (as Paul Churchland has noted),[60] then the belief that p is not identical to the belief that q if p is not identical to q. But the principles of propositional identity are tied to sentence identity in a language. In theory, of course, proposition identity can be specified in terms of sets of possible worlds, but in practice, the way such a set is referred to is as the set in which a particular (English) sentence is true. In reality, propositions are, for this reason, more like dollars than numbers, and the precision aspired to is an illusory goal.[61]

Lynne Rudder Baker gives a wonderful account of the reasons why the patterns discernible from the intentional stance should not be assumed to be repeated, somehow, in the brain. In this regard I especially commend her discussion of the bogus question about the location of the money making.[62] As she says, "Such questions are not serious spurs to inquiry." But other rather similar questions are. It is the confusion of the nonserious ones with the serious ones that causes a lot of the confusion. As Alan Turing noted, in one of his many prophetic asides,

> I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localize it.[63]

The reality of consciousness does not require its localization in the brain, but it still depends on features of brain activity, and if we want to confirm or disconfirm hypotheses about specific conscious experiences, what we need to test is the *truth* of the claims that constitute somebody's heterophenomenology. "It occurred to me that winters in Vermont are long" is not about the weather; and I might be wrong in asserting it. Nothing of the kind may have occurred to me. (Baker's analysis of this case would be much helped by honoring **David Rosenthal's** distinction between expressing and reporting: In uttering this sentence, I would be *expressing* a higher-order thought about my mental life in *reporting* a lower-order thought—about the weather.)

Heterophenomenology exhausts the intentional-stance theory of consciousness, but we want more (and so we should). Consider a parallel: There is undoubtedly a real pattern in the tales told (and believed) these days by self-styled victims of satanic ritual abuse, but are any of their beliefs true? We'd like to know. Similarly, there is certainly a real pattern in the tales told (and believed) by subjects about what occurs to them at various times, what they "do" in their minds at various times, and we'd like to know which of these beliefs of theirs are true. That is where "brain mapping" comes in. Baker sees a deep tension between the intentional stance and this brain-mapping move, mainly because she misinterprets me as thinking the brain

mapping will be a "deeper" theory (in her terms), and thus nonintentional. Not at all. The theory of content I espouse for the whole person I espouse all the way in. The neurobiological theory of content is homuncular functionalism, to dress it in its most vivid metaphorical costume, and hence the very same principles of interpretation are used to endow subpersonal parts with contents as are used to endow whole persons. (**David Rosenthal's** interpretive "hypothesis" on this score is thus correct. Since he has to work to arrive at this position, and Baker misses it, I cannot have done a proper job of expounding it.) The way in which personal-level attributions of belief and other intentional properties get confirmed (in the crunch) by subpersonal attributions of (nonordinary) intentional properties is roughly parallel to the way in which one might confirm one's attribution of culpable motives to, say, the British Empire, or the CIA, or IBM, by discovering a pattern of beliefs, desires, intentions, among the agents whose joint activity compose the actions, beliefs, and intentions of the superpersonal agent (see the discussion of **Carol Rovane**, at the end of this essay, for more on this).

So in the case of Eve, the story goes like this. Eve *expresses* the higher-order thought "I was suddenly conscious of the fact that I was not alone in the house," thereby *reporting* (truly or falsely) that she had a certain first-order thought to the effect that she was not alone in the house. Did she? We'll have to look at our record of what went on in her brain at the relevant time. Hmm, sure enough, here's a brain event that had the content (roughly): "Who's there!?!?!" That's close enough. We confirm her report, in this case. But we might not have confirmed it. We might have found circumstantial evidence to the effect that Eve has rationalized the whole event and wasn't the least bit driven, at the relevant time, by contents concerning the presence of others. Some of this evidence might be our secret videotape of her externally visible behavior at the time. (We see her humming contentedly throughout the relevant period, right up to the time when she answered the phone and began answering our questions about her recent phenomenology. There is no sign of apprehension, no abrupt change of trajectory, and the record of heart rate and skin conductance shows no alarm.) But we might actually get some even better (because closer to the gist of her self-report) evidence from our neurocryptography unit, which, applying the intentional stance to Eve's brain parts, has tentatively identified a homunculus whose duties include signalling a general fire drill whenever it detects the presence of another—it's hooked up to the vertical symmetry detector in the vision system, for instance.[64] That's the sort of story I had in mind. And it uses the intentional-stance theory of content all the way in.

At one point, Baker wonders what I would conclude if "the neuroscientist could find no brain state or process with which to identify the putatively conscious belief."[65] And she says either alternative would leave half of my project in the lurch. It's worse than that: Paraphrasing the physicists

who were confronted with the nonmagnet (if that's what it was), I'd have a much worse problem than which half of my project to cling to. I'd be worried that my whole project was on the verge of collapse. But she is right that in this terrible eventuality, I'd cling to the intentional stance aspect (which I *know* works very well) and do my best to explain why so many people have so many false beliefs about what is going on in their minds. Why would I be so eager to show that people *were mistaken* in these beliefs? Because my whole worldview resists as an utterly last resort any alternative explanation—e.g., a frankly dualistic explanation—of these undeniably real patterns. To pursue the parallel, suppose we can't confirm any of the peoples' stories about satanic cult abuse. No physical evidence of the events reported can be found. What would we do then? We would try to show that these folks, however sincere in their reports, were just wrong. But if the pattern defied that diagnosis, as it might, in principle, we'd have to start toying with the idea that the satanic ritual abuse *happened in another dimension* or some other equally extravagant departure from conservative physics and metaphysics. We're talking seriously spooky last resorts, folks!

Stephen Webb's dexterous navigation of the shoals of Plan One and the various Passes is largely for nought, since it misconstrues the intentional stance in a different way. It overlooks a key feature of the intentional stance: One is allowed to revise one's attributions in the light of falsified predictions. As soon as one hears the professor's surprising (in the circumstances) words, one revises the attribution: She must not have seen her keys, one easily concludes. So behavioral evidence has all along been front and center in the intentional stance. What is the heterophenomenological method, after all, but an application of this obvious principle? One patiently gathers behavioral evidence (largely but not entirely verbal behavioral evidence), hypothesizing interpretations and refining one's attributions until, in the limit, they account for ("predict," make sense of) all the behavior. The standard limit myth may be invoked: The beliefs and desires (and other intentional states) of an intentional system are those that would be attributed to it by an ideal observer with a God's-eye-point-of-view. Prediction and retrodiction are all one to the intentional stance, which explains as it attributes. It would not be a very useful stance if it couldn't be harnessed reliably for real-time prediction, but after-the-fact explanation of behavior is hardly off limits to it, and the behavior itself is obviously a main source of evidence.

Webb patiently tries to corral me at various other choice points, overlooking reasons I have already given for resisting his alternatives. Much of what I have said above in response to **McLaughlin** and **O'Leary-Hawthorne**, and **Baker**, applies here as well. To reiterate: The skin is not that important as a boundary, as Skinner famously conceded (with my concurrence), so "internal" behavior is in principle not off limits to the interpretive exercise, and it can, in particular cases, crucially supplement the

stock of relevant data, but it gets interpreted by the same rules as behavior outside the skin: The carton of cigarettes listed on the scrap of paper has the same evidentiary status as the carton of cigarettes referred to by some cerebral shopping list, if and when science can discern it. I have pointed out that the line between the design stance and the intentional stance is not sharp in any case; subpersonal cognitive psychology is a design stance enterprise conducted with the aid of liberal intentional stance characterizations of homunculi.

IV. SUPERFICIALISM VERSUS HYSTERICAL REALISM

> σώζειν τὰ φαινόμενα [Save the phenomena.]
> —Plato
>
> Save the surface and you save all.
> —Sherwin-Williams

Thus Quine opens his *Pursuit of Truth*.[66] **Georges Rey** is surely right that the key to the profound disagreements he and other Fodorians have with me is that I espouse a view he deftly labels "superficialism." Who would ever want to be called a superficialist? Well, Quine, for one, might not shrink from the label. One good term deserves another. In the past I've called Fodor an "industrial strength realist," but the connotations of that term are all too positive, and I think Rey's blithe self-description as a "common-sense" realist is amusingly belied by the convolutions of doctrine he presents to us, so I'll change the epithet. Rey is an advocate of *hysterical realism*. Now you can choose: Which would you rather be, a hysterical realist or a superficialist like me?

Rey is far from being alone when he responds to the siren song of hysterical realism. The defense of what Richard Rorty mockingly calls "Our Realist Intuitions" is a fervent activity in many quarters, and while I am not hereby endorsing Rorty's brand of resistance, I do think he has touched one of the untouchable sore spots in contemporary philosophy. Hysterical realism[67] deserves its name, I will argue, because it is an overreaction, a rationally unmotivatable spasm brought on by peering into the abyss of certain indeterminacies that really should not trouble anybody so much.

For Rey the particular topic is consciousness, a contentious topic, but we can begin cautiously. I trust we can all agree on this much: We find people asking questions about consciousness, using ordinary language, of course, with its ordinary presuppositions enshrined. Then once we start examining the real, causal complexities of what happens in brains, we find, I claim, a strikingly poor match between the pretheoretical presuppositions and the messy details. Here is another place where a disagreement over empirical facts looms large. Those who have not looked very closely are apt to think